Research Report

Joint and Conditional Estimation for Implicit Models for Tests With Polytomous Item Scores

Shelby J. Haberman

Research & Development

March 2006
RR-06-03

Joint and Conditional Estimation for Implicit Models for Tests With Polytomous Item Scores

Shelby J. Haberman
ETS, Princeton, NJ

March 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Copyright © 2006 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and TOEFL are registered trademarks of Educational Testing Service (ETS).

SAT is a registered trademark of the College Board. SAT Reasoning Test is a trademark of the College Board.

Abstract

Multinomial-response models are available that correspond implicitly to tests in which a total score is computed as the sum of polytomous item scores. For these models, joint and conditional estimation may be considered in much the same way as for the Rasch model for right-scored tests. As in the Rasch model, joint estimation is only attractive if both the number of items and the number of examinees are large, while conditional estimation can be employed for a large number of examinees whether or not the number of items is large. In neither case is computation difficult given currently available computers. Large-sample results favor use of conditional estimation, although some use of joint estimation can be contemplated if the number of items is large.

Key words: Maximum likelihood, consistency, asymptotic normality

Acknowledgements

The author would like to thank Matthias von Davier, Neil Dorans, and Paul Holland for helpful discussions.

Introduction

Probability models based on exponential families are readily constructed for tests in which the total score is the sum of polytomous item scores. In these models, which are described in Section 1., the total scores for the examinees are part of the sufficient statistics for the model. These models can be employed to assess what information, if any, can be obtained concerning examinees that is not revealed by the total score. In addition, these models typically have parameters that correspond to such common concepts in item response theory as examinee ability and item difficulty. The models themselves have existed for some time (Bock, 1972; Andrich, 1978; Masters, 1982; Andersen, 1983); however, the appropriateness of joint and conditional estimation has not been extensively studied for these models in the common case in which both the number of examinees and the number of items are large. In this report, two aspects of joint and conditional estimation are considered for this case. Consistency and asymptotic normality of parameter estimates are explored for these methods, and computation of estimates is considered. Marginal estimation for this class of models is not considered in this report, and this topic does merit study. This report confines attention to techniques that do not require assumptions about the underlying ability distribution.

Joint and conditional estimation proceed in much the same way as for the Rasch model for binary responses (Rasch, 1960; Haberman, 1977, 2004). As in the Rasch model for binary responses, straightforward application of maximum likelihood presents a number of complications if no restrictions are imposed on the ability distribution, so that joint maximum likelihood and conditional maximum likelihood will receive considerable attention.

Section 2. examines joint maximum-likelihood estimation (JMLE). Results rely heavily on previously derived results for the binary Rasch model (Andersen, 1972; Fischer, 1981; Haberman, 1977, 2004). As expected, JMLE does not lead to fully satisfactory approximate confidence intervals for item parameters, and the normal approximation for the distribution of ability estimates is not fully satisfactory. Nonetheless, joint estimation does have possible use in construction of starting values for conditional estimation.

Section 3. examines conditional maximum-likelihood estimation (CMLE) for the

models under study. Techniques are based on those for the binary Rasch model (Andersen, 1972, 1973a, 1973b; Fischer, 1981; Haberman, 2004). Basic properties of conditional maximum-likelihood estimates are readily examined. Computation with the Newton-Raphson algorithm is only moderately more complicated than for the binary Rasch model provided that convolutions are used and starting values exploit joint estimation. Normal approximations for estimates of item parameters are established that apply whether or not the number of items increases.

To illustrate results, data from Form A of the TOEFL® field trial are used. To compare estimates, the reading and listening sections are examined as a single test. Although the preponderance of items have simple right scores, one reading item has integer scores from 0 to 4, one has integer scores 0 to 2, and one has integer scores 0 to 3. Two listening items have integer scores 0 to 2. In all, 71 items are scored on 2,720 examinees. Use of the single test provides a better opportunity for joint estimation than is afforded by a separate reading and a separate listening tests. In addition, it is easily verified that the listening and reading test results are very highly correlated, so that it is not obvious that much error is introduced by combining the scales. The actual loss of information from this step is to be considered in a separate paper.

Section 4. summarizes the implications of the research for psychometric practice and discusses some further areas of possible development.

1. Models for Polytomous Scoring

In the basic model for polytomous scoring, the number of possible scores per item may vary, but it is required that all scores be rational numbers. This requirement is necessary for conditional estimation. The model examined is the nominal model (Bock, 1972; Andersen, 1983). The rating scale model (Andrich, 1978) and partial credit model (Masters, 1982) are considered in relationship to the basic model.

In the model under study, the score of examinee i, $1 \leq i \leq n$, is a sum of scores assigned to each of q items. For each item j and examinee i, let the response be denoted by the integer Y_{ij}. For some integer $r_j \geq 2$, let $0 \leq Y_{ij} \leq r_j - 1$, and for each Y_{ij}, let the possible scores on item j be $u_j(k)$, $0 \leq k \leq r_j - 1$. Let the score of examinee i on item j be

$u_{ij} = u_j(Y_{ij})$, so that the total examinee score is

$$S_i = \sum_{j=1}^{q} u_{ij}.$$

In the TOEFL case with reading and listening scores combined, $q = 71$ and $r_j = 2$ except for Items 11, 25, 38, 42, and 58. For Item 11, $r_{11} = 5$; for Item 38, $r_{38} = 4$; and for Items j equal to 25, 42, and 58, $r_j = 3$. In the TOEFL example, $u_j(k)$ is always k. In the math and verbal sections of the SAT® I examination, which was recently replaced with the SAT Reasoning Test™, a somewhat more complex system of scoring is used. If item j is a multiple-choice item with $d_j > 1$ alternatives, then a score of 1 is used for a correct answer, a score of $-1/(d_j - 1)$ is used for an incorrect answer, and a score of 0 is used for an omitted response. In grid-in responses, a score of 1 is used for a correct response. No response or an incorrect response receives a score of 0. It is easily seen that the SAT scoring method is a special case of the scoring method considered in this paper. Unlike the TOEFL example, the scores for items are not necessarily integers and are not necessarily nonnegative.

It is assumed that the vectors $\mathbf{Y}_i$ with coordinates Y_{ij}, $1 \leq j \leq q$, are mutually independent and identically distributed. It is also assumed that, for each item j, the possible scores $u_j(k)$ are not all equal, so that the item response can change the total score. If $\hat{u}_j$ is the arithmetic mean

$$\bar{u}_j = \frac{1}{r_j} \sum_{k=1}^{r_j} u_j(k),$$

then

$$U_j = \sum_{k=1}^{r_j} [u_j(k) - \bar{u}_j]^2 > 0.$$

Because conditional estimation is often considered, the added assumption is made that each score $u_j(k)$ is equal to $u_{jn}(k)/u_{jd}(k)$ for an integer $u_{jn}(k)$ and a positive integer $u_{jd}(k)$, so that $u_j(k)$ is a rational number. This requirement is needed to permit useful inferences conditional on the examinee scores S_i.

In the nominal model, to each examinee i corresponds an unknown ability parameter θ_i, and the θ_i are independent random variables with common unknown distribution function D. Given θ_i, the Y_{ij}, $1 \leq j \leq q$, and the θ_h, $h \neq i$, are mutually independent. To each item j correspond unknown item parameters β_{jk}, $0 \leq k \leq r_j - 1$. To construct vectors to

use with these parameters, let $R_0 = 0$ and $R_j = R_{j-1} + r_j$ for $1 \leq j \leq q$. Let $\boldsymbol{\beta}$ be the vector of dimension R_q with coordinate $\zeta(j,k) = R_{j-1} + k + 1$ equal to β_{jk}, $1 \leq j \leq q$ and $0 \leq k \leq r_j - 1$. Under the nominal model, the conditional probability that $Y_{ij} = k$ given θ_i is

$$p_{ijk} = p_{jk}(\boldsymbol{\beta}, \theta_i), \tag{1}$$

where

$$e_j(\boldsymbol{\beta}, \theta_i) = \left\{ \sum_{k=0}^{r_j - 1} \exp[\theta_i u_j(k) - \beta_{jk}] \right\}^{-1} \tag{2}$$

and

$$p_{jk}(\boldsymbol{\beta}, \theta_i) = e_j(\boldsymbol{\beta}, \theta_i) \exp[\theta_i u_j(k) - \beta_{jk}] \tag{3}$$

(Bock, 1972; Andersen, 1983). To permit identification of parameters, the convention is adopted that $\boldsymbol{\beta}$ is in the set Λ of R_q-dimensional vectors x with coordinate $\zeta(j,k)$ equal to x_{jk} such that $\sum_{k=0}^{r_j - 1} x_{jk} = 0$ for each item j and $\sum_{k=0}^{r_1 - 1} [u_1(k) - \bar{u}_1] x_{1k} = 0$.

Conditional on the θ_i, sufficient statistics for the observed Y_{ij}, $1 \leq i \leq n$, $1 \leq j \leq q$, are the examinee scores S_i and the number of examinees f_{jk} with $Y_{ij} = k$, $0 \leq k \leq r_j - 1$, $1 \leq j \leq q$, and $1 \leq i \leq n$. The nominal model is the model implicitly defined by the requirement of sufficiency of the S_i and the f_{jk} given the θ_i (Gilula & Haberman, 2000).

Special cases of the nominal model for polytomous item scores can be found in the literature. The Rasch model for binary data arises if $r_j = 2$ and $u_j(k) = k$ for each j (Rasch, 1960). In this case, the identifiability restrictions are equivalent to the requirements that $\beta_{j1} = -\beta_{j0}$ for $j \geq 2$ and $\beta_{10} = \beta_{11} = 0$. In the partial credit model, $u_j(k) = k$ and r_j is a constant r, so that $\sum_{k=0}^{r-1} \beta_{jk} = 0$ and $\sum_{k=0}^{r-1} [k - (r-1)/2] \beta_{1k} = 0$ (Masters, 1982). In a version of the rating scale model, r_j is a constant r, $u_j(k)$ is independent of j, and

$$\beta_{jk} = \mu_k - \nu_j [u_1(k) - \bar{u}_1]$$

for unknown μ_k and ν_j such that $\sum_{k=0}^{r-1} \mu_k = 0$ and $\sum_{k=0}^{r-1} [u_1(k) - \bar{u}_1] \mu_k = \nu_1 U_1$. The ν_j are item difficulties. The conditions on β_{jk} are satisfied if $\mu_r = 0$, $u_1(r) = 0$, and

$$\nu_1 = U_j^{-1} \sum_{k=1}^{r} [u_1(k) - \bar{u}_1] \mu_k$$

(Andrich, 1978).

For all versions of the nominal model, the probability that $\mathbf{Y}_i$ has a specific value $\mathbf{c}$ is readily calculated. Consequently, a log likelihood function can be obtained. To calculate the desired probability, let Γ be the set of q-dimensional vectors $\mathbf{c}$ with integer coordinates c_j, $1 \le j \le q$, such that $1 \le c_j \le r_j$. Then for $\mathbf{c}$ in Γ, the probability $p_J(\mathbf{c})$ that $\mathbf{Y}_i = \mathbf{c}$ is

$$p_J(\mathbf{c}) = E\left(\prod_{j=1}^{n} p_{1jc_j}\right).$$

For a more explicit expression, let

$$S(\mathbf{c}) = \sum_{j=1}^{q} u_j(c_j)$$

be the score $S_i = S(\mathbf{c})$ obtained if $\mathbf{Y}_i = \mathbf{c}$. Let $\mathcal{S}$ be the set of possible values of S_i, so that s is in $\mathcal{S}$ if, and only if, $s = S(\mathbf{c})$ for some $\mathbf{c}$ in Γ. Let A be the number of elements of $\mathcal{S}$, and let $s(a)$ be the ath smallest element of $\mathcal{S}$ for a from 1 to A. For s in $\mathcal{S}$, let $\Gamma(s)$ be the set of $\mathbf{c}$ in Γ such that $S(\mathbf{c}) = s$. For R_q-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ with respective coordinates $\zeta(j,k)$ equal to x_{jk} and y_{jk}, $1 \le j \le q$ and $0 \le k \le r_j - 1$, let

$$\mathbf{x}'\mathbf{y} = \sum_{j=1}^{q} \sum_{k=0}^{r_j-1} x_{jk} y_{jk}.$$

For any $\mathbf{c}$ in Γ, let $Z_{jk}(\mathbf{c})$ be 1 for $c_j = k$ and 0 otherwise, and let $\mathbf{Z}(\mathbf{c})$ denote the R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to $Z_{jk}(\mathbf{c})$, $1 \le j \le q$ and $0 \le k \le r_j - 1$. Then

$$\sum_{j=1}^{q} \beta_{jc_j} = \sum_{j=1}^{q} \sum_{k=0}^{r_j-1} Z_{jk}(\mathbf{c}) \beta_{jk} = \boldsymbol{\beta}'\mathbf{c}.$$

Let

$$M_s(\boldsymbol{\beta}) = \sum_{\mathbf{c} \in \Gamma(s)} \exp[-\boldsymbol{\beta}'\mathbf{Z}(\mathbf{c})] \tag{4}$$

for s in $\mathcal{S}$, let

$$\Phi(\boldsymbol{\beta}, \theta) = \sum_{s \in \mathcal{S}} \exp(\theta s) M_s(\boldsymbol{\beta}),$$

and let

$$\tau_s = \log \int [\Phi(\boldsymbol{\beta}, \theta)]^{-1} \exp(\theta s) dD(\theta). \tag{5}$$

Let $\boldsymbol{\tau}$ be the A-dimensional vector with coordinate a equal to $\tau_{s(a)}$ for a from 1 to A. For any $\mathbf{c}$ in $\Gamma(s)$ and any s in $\mathcal{S}$,

$$p_J(\mathbf{c}) = \exp[-\boldsymbol{\beta}'\mathbf{Z}(\mathbf{c}) + \tau_s]. \tag{6}$$

Let $\mathbf{p}_J$ be the array of $p_J(\mathbf{c})$ for $\mathbf{c}$ in Γ. Let Ξ consist of all $\mathbf{p}_J$ such that (6) holds for $\mathbf{c}$ in $\Gamma(s)$ and s in $\mathcal{S}$, where τ_s satisfies (5) for s in $\mathcal{S}$ for some $\boldsymbol{\beta}$ in Λ and some distribution function D. To obtain the log likelihood function $\ell(\mathbf{p}_J)$, let Z_{ijk} be 1 if $Y_{ij} = k$ and let $Z_{ijk} = 0$ if $Y_{ij} \neq k$, and let

$$Z_{+jk} = \sum_{i=1}^{n} Z_{ijk}$$

be the number of examinees i who provide response k to item j. Let $\mathbf{Z}_+$ be the R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to Z_{+jk}, $1 \leq j \leq q$ and $0 \leq k \leq r_j - 1$. Let $N_S(s)$ be the number of examinees i with total score $S_i = s$, and let $\mathbf{N}_S$ be the array of $N_S(s)$ for s in $\mathcal{S}$. Let $\mathbf{N}_S$ be the vector of $N_S(s(a))$, a from 1 to A, and let

$$\mathbf{N}_S'\boldsymbol{\tau} = \sum_{s\in\mathcal{S}} N_S(s)\tau_s.$$

Let $\mathbf{p}_J$ be the array of $p_J(\mathbf{c})$ for $\mathbf{c}$ in Γ. Then

$$\begin{aligned} \ell(\mathbf{p}_J) &= \sum_{i=1}^{n} \log p_J(\mathbf{Y}_i) \\ &= -\boldsymbol{\beta}'\mathbf{Z}_+ + \mathbf{N}_S'\tau. \end{aligned}$$

Thus $\mathbf{Z}_+$ and $\mathbf{N}_S$ are jointly sufficient for $\boldsymbol{\beta}$ and D.

Use of maximum likelihood with the nominal model is far from straightforward due to the integrals involved in the definition of τ_s and due to the lack of identifiability of the distribution function D. The problem is similar to difficulties encountered with the Rasch model (Cressie & Holland, 1983; Haberman, 2004). The probability $p_S(s)$ that $S_i = s$ is

$$p_S(s) = M_s(\boldsymbol{\beta})\exp(\tau_s),$$

so that

$$\sum_{s\in\mathcal{S}} M_s(\boldsymbol{\beta})\exp(\tau_s) = \sum_{s\in\mathcal{S}} p_S(s) = 1.$$

Because each $u_j(k)$ is rational, a largest positive rational number B exists such that any member s of $\mathcal{S}$ is $s(1) + hB$ for a nonnegative integer $h \leq [s(A) - s(1)]/B$. In the TOEFL example under study, B is 1. If $s = s(1) + hB$ for an integer h and s is in $\mathcal{S}$, then $\exp(\tau_s - \tau_{s(1)})$ is the hth moment of a positive random variable X such that the probability that $\log X \leq y$, y real, is

$$\frac{\int_{-\infty}^{y} [\Phi(\boldsymbol{\beta}, \theta)]^{-1} \exp[s(1)\theta] dD(\theta)}{\int_{-\infty}^{\infty} [\Phi(\boldsymbol{\beta}, \theta)]^{-1} \exp[s(1)\theta] dD(\theta)}.$$

Because only a finite number of moments are specified by the ratios $\exp(\tau_s - \tau_{s(1)})$, it follows that more than one distribution function D corresponds to the same $\boldsymbol{\beta}$ and τ_s, s in $\mathcal{S}$. On the other hand, if a positive random variable X exists such that $\exp(\tau_s - \tau_{s(1)})$ is the hth moment of X whenever $s = s(1) + hB$ is in $\mathcal{S}$ and if G is the distribution function of $\log X$, then (5) holds if

$$D(x) = \frac{\int_{-\infty}^{x} \Phi(\boldsymbol{\beta}, \theta) \exp(-s(1)\theta) dG(\theta)}{\int_{-\infty}^{\infty} \Phi(\boldsymbol{\beta}, \theta) \exp(-s(1)\theta) dG(\theta)}.$$

The nominal model implies the log-linear model in which, for some $\boldsymbol{\beta}$ in Λ, and real τ_s, s in $\mathcal{S}$,

$$\log p_J(\mathbf{c}) = -\boldsymbol{\beta}' \mathbf{Z}(\mathbf{c}) + \tau_s \tag{7}$$

for $\mathbf{c}$ in $\Gamma(s)$ and s in $\mathcal{S}$ and

$$\sum_{s \in \mathcal{S}} M_s(\boldsymbol{\beta}) \exp(\tau_s) = 1. \tag{8}$$

Let Ξ_+ consist of all $\mathbf{p}_J$ such that (7) holds for some $\boldsymbol{\beta}$ in Λ and some τ_s, s in $\mathcal{S}$, such that (8) holds. Then $\mathbf{p}_J$ satisfies the log-linear extension of the nominal model if, and only if, $\mathbf{p}_J$ is in Ξ_+. On the other hand, the log-linear model does not imply the nominal model, for the nominal model can only hold if τ_s satisfies the convexity condition that

$$\tau_s \leq a\tau_t + (1 - a)\tau_u$$

whenever $s = at + (1 - a)u$, s, t, and u are in $\mathcal{S}$, and $0 < a < 1$ (Feller, 1966, p. 153). Thus Ξ is a proper subset of Ξ_+. For the log-linear extension of the nominal model, the log-likelihood $\ell(\mathbf{p}_J)$ has the same form as in the nominal model, and the sufficient statistics remain $\mathbf{Z}_+$ and $\mathbf{N}_S$.

2. Joint Maximum-Likelihood Estimation

In JMLE, the ability parameters θ_i are regarded as fixed parameters to be estimated. The estimates of the θ_i are then used to estimate the distribution function D. Use of joint maximum likelihood has a long history of controversy in many areas of statistics (Kiefer & Wolfowitz, 1956). In many circumstances, joint maximum likelihood is relatively easily implemented; however, consistency of estimates is a major concern, especially if the number of items is fixed and the number of subjects increases. Consistency issues can be resolved if both the number of items and the number of subjects increases, a result that is known in the special case of the binary Rasch model (Haberman, 1977, 2004). In this section, it is shown that joint estimation is rather unsatisfactory in terms of consistency if the number of items is not large, but joint estimation can lead to consistent and asymptotically normal parameter estimates if both the number of items and the number of examinees is large.

To simplify large-sample results, a number of boundedness assumptions and convergence assumptions are made. To begin, it is assumed that the θ_i are bounded, so that $D(x)$ is 0 for x sufficiently small, and $D(x) = 1$ for x sufficiently large. It is also assumed that the β_{jk}, $u_{jn}(k)$, $u_{jd}(k)$, and r_j are uniformly bounded if q goes to ∞, so that B has the same value for all q sufficiently large. Let $r_{\max}$ be the largest value of r_j for any $j \geq 1$. It is assumed that, for each integer $r \leq r_{\max}$, the fraction of items j with $r_j = r$ approaches a constant f_j as q increases and the empirical distribution of $\boldsymbol{\beta}_j$, $r_j = r$, converges weakly to the distribution of the r-dimensional random vector $\boldsymbol{\beta}_r^*$. The assumptions made imply that constants s_-^* and s_+^* exist such that $s(1)/q$ converges to s_-^* and $s(A)/q$ converges to s_+^*.

To define joint estimation, let $\mathbf{p}$ denote the array of p_{ijk}, $1 \leq i \leq n$, $1 \leq j \leq q$, $1 \leq k \leq r_j$. The joint log likelihood function

$$\ell_J(\mathbf{p}) = \sum_{i=1}^{n} \sum_{k=1}^{q} Z_{ijk} \log p_{ijk}$$

is maximized under the model constraints. In the expression for $\ell_J(\mathbf{p})$, note that $Z_{ijk} \log p_{ijk}$ is $\log p_{ijh}$ if $Y_{ij} = h$. The resulting maximum ℓ_{JM} under the constraints from (1) is achieved if, and only if, $\beta_{jk} = \hat{\beta}_{jk}$ and $\theta_i = \hat{\theta}_i$ for $\hat{\beta}_{jk}$ and $\hat{\theta}_i$ such that $\hat{\boldsymbol{\beta}}$ is the R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to $\hat{\beta}_{jk}$,

$$\hat{p}_{ijk} = p_{jk}(\hat{\boldsymbol{\beta}}, \hat{\theta}_i), \tag{9}$$

$\hat{\boldsymbol{\beta}}$ is in Λ,

$$\sum_{j=1}^{q}\sum_{k=0}^{r_j-1} u_j(k)\hat{p}_{ijk} = S_i \tag{10}$$

and

$$Z_{+jk} = \hat{p}_{+jk} = \sum_{i=1}^{n} \hat{p}_{ijk} \tag{11}$$

(Haberman, 1977). If the $\hat{\beta}_{jk}$ and $\hat{\theta}_i$ exist, then they are uniquely defined. The $\hat{\beta}_{jk}$ are the JMLEs of the β_{jk}, and the $\hat{\theta}_i$ are the JMLEs of the θ_i. The vector $\hat{\boldsymbol{\beta}}$ is the maximum-likelihood estimate of $\boldsymbol{\beta}$.

2.1 Computations and Collapsed Tables

Computation of JMLEs is greatly simplified by use of a collapsed table. Let $\mathcal{S}_+$ be the set of s in $\mathcal{S}$ such that $N_S(s) > 0$. Consider the array with entries f_{sjk} for s in $\mathcal{S}_+$, $0 \le k \le r_j - 1$, and $1 \le j \le q$, such that f_{sjk} is the number of examinees i, $1 \le i \le n$, such that $S_i = s$ and $Y_{ij} = k$. For real x and y, let $\delta_x(y)$ be 1 for $x = y$ and 0 otherwise. Observe that

$$\begin{aligned}
\sum_{j=1}^{q}\sum_{k=0}^{r_j-1} u_j(k) f_{sjk} &= \sum_{j=1}^{q}\sum_{k=0}^{r_j-1} u_j(k) \sum_{i=1}^{n} \delta_s(S_i) u_j(k) Z_{ijk} \\
&= \sum_{i=1}^{n} \delta_s(S_i) \sum_{j=1}^{q}\sum_{k=0}^{r_j-1} u_j(k) Z_{ijk} \\
&= s N_S(s)
\end{aligned}$$

for s in $\mathcal{S}_+$, and the sum

$$\begin{aligned}
f_{+jk} &= \sum_{s\in\mathcal{S}_+} f_{sjk} \\
&= \sum_{s\in\mathcal{S}_+}\sum_{i=1}^{n} \delta_s(S_i) Z_{ijk} \\
&= \sum_{i=1}^{n} Z_{ijk} \sum_{s\in\mathcal{S}_+} \delta_s(S_i) \\
&= Z_{+jk}.
\end{aligned}$$

Consider maximization of the collapsed log likelihood

$$\ell_{JC}(\mathbf{p}_C) = \sum_{s \in \mathcal{S}_+} \sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} f_{sjk} \log p_{sjkC},$$

for $\mathbf{p}_C$ the array of p_{sjkC}, s in $\mathcal{S}_+$, $0 \leq k \leq r_j - 1$, $1 \leq j \leq q$, with the constraints that

$$p_{sjkC} = p_{jk}(\boldsymbol{\beta}, \theta_{sC})$$

and $\boldsymbol{\beta}$ is in Λ. Let ℓ_{JCM} be the supremum of ℓ_{JC}. Then $\ell_{JCM} = \ell_{JC}$ if, and only if, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_C$ and $\theta_{sC} = \hat{\theta}_{sC}$, where $\hat{\boldsymbol{\beta}}_C$ is in Λ,

$$\hat{p}_{sjkC} = p_{jk}(\hat{\boldsymbol{\beta}}_C, \hat{\theta}_{sC}) \tag{12}$$

for s in $\mathcal{S}_+$, $0 \leq k \leq r_j - 1$, and $1 \leq j \leq q$,

$$\sum_{s \in \mathcal{S}_+} N_S(s) \hat{p}_{sjkC} = f_{+jk} = Z_{+jk} \tag{13}$$

for $0 \leq k \leq r_j - 1$ and $1 \leq j \leq q$, and

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} u_j(k) \hat{p}_{sjkC} = s \tag{14}$$

for s in $\mathcal{S}_+$ (Haberman, 1977, 2004). The vector $\hat{\boldsymbol{\beta}}_C$ is uniquely defined if it exists. The $\hat{\theta}_{sC}$ are uniquely defined for s in $\mathcal{S}_+$ if they exist.

Because $\hat{p}_{sjkC}$ is positive and less than 1 for s in $\mathcal{S}_+$, $0 \leq k \leq r_j - 1$, and $1 \leq j \leq q$, (14) does not hold if $s(1)$ or $s(A)$ is in $\mathcal{S}_+$, so that some examinee i exists such that either

$$u_j(Y_{ij}) = u_{j+} = \max_{0 \leq k \leq r_j - 1} u_j(k)$$

for all items j or

$$u_j(Y_{ij}) = u_{j-} = \min_{0 \leq k \leq r_j - 1} u_j(k)$$

for all items j. In the case of the TOEFL examination, $A = 80$, $s(1) = 0$, and $s(A) = 79$. One examinee achieved a total score of 79, so nonexistence is an issue.

The relationship of $\hat{\theta}_{sC}$ and $\hat{\boldsymbol{\beta}}_C$ to the corresponding joint maximum-likelihood estimates is straightforward. If $\hat{\theta}_{sC}$ and $\hat{\boldsymbol{\beta}}_C$ exist, then $\hat{\theta}_i = \hat{\theta}_{sC}$ for $S_i = s$ and $\hat{\beta}_{jk} = \hat{\beta}_{jkC}$. If the

$\hat{\beta}_{jk}$ and $\hat{\theta}_i$ exist, then $\hat{\beta}_{jkC} = \hat{\beta}_{jk}$ and $\hat{\theta}_{sC} = \hat{\theta}_i$ for $s = S_i$. Thus joint maximum-likelihood estimates are readily found by maximization of ℓ_{JC}.

From a computational standpoint, the collapsed table has major impact, for one can compute joint maximum-likelihood estimates by acting as if a multinomial response model holds with independent arrays f_{sjk}, $0 \le k \le r_j - 1$, with sample size $N_S(s)$ and with probabilities p_{sjk}, $0 \le k \le r_j - 1$, for $1 \le j \le q$ and s in $\mathcal{S}_+$, where

$$p_{sjk} = p_{jk}(\boldsymbol{\beta}, \theta_{sC}).$$

Instead of an n by q array of responses Y_{ij}, it suffices to consider the array of counts f_{sjk}. If $r_+ = \sum_{j=1}^{q} r_j$, then the array of f_{sjk} has no more than Ar_+ elements. For instance, in the TOEFL example, the array has no more than $80 \times 150 = 12,000$ entries, but there are $2,720 \times 71 = 193,120$ responses Y_{ij} to consider. The array of f_{sjk} also assists in the study of existence of joint maximum-likelihood estimates and in the study of large-sample properties of JMLE. Given existing software for multinomial response models, computation of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_C$ and $\hat{\theta}_{sC}$, s in $\mathcal{S}_+$, is straightforward.

2.2 Existence of Joint Maximum-Likelihood Estimates

Existence of joint maximum-likelihood estimates is a substantial problem in practice. To study the issue, standard results from the theory of log-linear models are used as in the following theorem (Haberman, 1974, chap. 2):

Theorem 1 *Joint maximum-likelihood estimates exist if, and only if, a table of positive g_{sjk}, $0 \le k \le r_j - 1$, $1 \le j \le q$, s in $\mathcal{S}_+$, exists such that $g_{+jk} = f_{+jk}$ for $0 \le k \le r_j - 1$ and $1 \le j \le q$, $\sum_{k=0}^{r_j - 1} g_{sjk} = N_S(s)$ for $1 \le j \le q$ and s in $\mathcal{S}_+$, and*

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j} u_j(k) g_{sjk} = s N_S(s)$$

for s in $\mathcal{S}$.

It is clearly true that joint maximum-likelihood estimates exist if f_{sjk} is positive for all s in $\mathcal{S}_+$, $0 \le k \le r_j - 1$, and $1 \le j \le q$, for one may just take $g_{sjk} = f_{sjk}$. It is clearly true

that joint maximum-likelihood estimates do not exist if f_{+jk} is 0 for some j and k or if $s(1)$ or $s(A)$ is in $\mathcal{S}_+$.

These results suffice to indicate that joint maximum-likelihood estimates do not exist for the TOEFL example, for $N_S(s(A)) = 1 > 0$. Thus a more general approach to joint estimation is required for the TOEFL data.

2.3 Extended Joint Maximum-Likelihood Estimates

Without any conditions, extended joint maximum-likelihood estimates $\hat{p}_{ijk}$ of p_{ijk} may be defined such that $0 \leq \hat{p}_{ijk} \leq 1$, $\hat{p}_{ij+} = 1$, $\hat{p}_{+jk} = Z_{+jk}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} Z_{ijk} \log \hat{p}_{ijk} = \ell_{JM},$$

and real $\theta_{i\nu}$, $1 \leq i \leq n$, and $\boldsymbol{\beta}_\nu$ in Λ exist for $\nu \geq 0$ such that

$$p_{jk}(\boldsymbol{\beta}_\nu, \theta_{i\nu})$$

approaches $\hat{p}_{ij}$ as ν approaches ∞ (Haberman, 1974, pp. 402–404). The definition of extended joint maximum-likelihood estimates $\hat{p}_{ijk}$ is consistent with the previous definition of $\hat{p}_{ijk}$ when joint maximum-likelihood estimates exist. The $\hat{p}_{ijk}$ are uniquely defined. In addition, if real $\theta_{i0\nu}$ and $\boldsymbol{\beta}_{0\nu}$ in Λ exist for $1 \leq i \leq n$ and $\nu \geq 0$ such that

$$p_{jk}(\boldsymbol{\beta}_{0\nu}, \theta_{i0\nu})$$

approaches $\hat{p}_{ijk0}$ as ν approaches ∞,

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} u_j(k) p_{ijk0} = S_i,$$

and $p_{+jk0} = Z_{+jk}$, then $p_{ijk0} = \hat{p}_{ijk}$. In terms of the collapsed table, $\hat{p}_{sjkC}$ are defined so that (13) and (14) hold, and $\hat{p}_{ijk} = \hat{p}_{sjkC}$ for $S_i = s$. In this case,

$$\sum_{s \in \mathcal{S}_+} \sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} f_{sjk} \log \hat{p}_{sjkC} = \ell_{JCM}.$$

The estimates $\hat{p}_{sjkC}$ may be used to create estimates $\hat{\theta}_i$, $\hat{\theta}_{sC}$, and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_C$. These estimates may be infinite in some cases. For instance, $\hat{\theta}_i = \hat{\theta}_{sC} = \infty$ if $S_i = s(A)$, and

$\hat{\theta}_i = \hat{\theta}_{sC} = -\infty$ if $S_i = s(1)$. If the $\hat{p}_{s1kC}$ are positive for all k from 1 to r_1 and $S_i = s$, then

$$\hat{\theta}_i = \hat{\theta}_{sC} = U_1^{-1} \sum_{k=0}^{r_1-1} [u_1(k) - \bar{u}_1] \log \hat{p}_{s1k}.$$

If, in addition, the $\hat{p}_{sjk}$ are positive for $0 \leq k \leq r_j - 1$ for some $j \geq 1$, then coordinate $\zeta(j,k)$ of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\beta}_{jk} = -\log(\hat{p}_{sjkC} + r_j^{-1} \sum_{k'=0}^{r_j-1} \hat{p}_{sjk'C}) + \hat{\theta}_s[u_j(k) - \bar{u}_j].$$

2.4 Consistency

Even if (1) holds, if the number q of items is constant, the β_{jk} are constant, and n approaches ∞, then the $\hat{\theta}_i$ are not consistent estimates of the θ_i, and the $\hat{\beta}_{jk}$ are not consistent estimates of the β_{jk}. This outcome is predictable given results for the Rasch model for binary data (Andersen, 1973a, pp. 66–69). Indeed, the probability approaches 1 that ordinary maximum-likelihood estimates do not even exist, a result expected given similar results for the binary Rasch model (Haberman, 1977).

A much more subtle problem arises if the number q of items increases as the number n of examinees increases. Given previous results for the binary Rasch model (Haberman, 1977, 2004), it is reasonable to expect that consistency results would be available in this case. As shown in this section, this expectation is indeed fulfilled. One finds that $\max_{1 \leq j \leq q} \max_{0 \leq k \leq r_j - 1} |\hat{\beta}_{jk} - \beta_{jk}|$ converges in probability to 0, and, for any given examinee i, $\hat{\theta}_i - \theta_i$ converges in probability to 0. It also follows that the empirical distribution function $\hat{D}$ of the $\hat{\theta}_i$ converges weakly with a probability of 1 to the distribution function D of θ_1.

A fixed number of items. For a fixed number of items, the existence issue is quite straightforward for ordinary joint maximum likelihood, although consistency requires a more careful argument. Consider the following theorems.

Theorem 2 *Let the number q of items be fixed, and let the number n of examinees approach ∞. Then the probability that joint maximum-likelihood estimates exist approaches 0.*

Proof. Let P_s, s in $\mathcal{S}$, be the unconditional probability $P(S_i = s)$ that $S_i = s$. Then each P_s is positive, so that the probability is positive that examinee i has either a minimal

formula score $S_i = s(1)$ or a maximal formula score $s(A)$. Joint maximum-likelihood estimates only can exist if $s(1) < S_i < s(A)$ for each examinee i from 1 to n. The probability that $s(1) < S_i < s(A)$ for $1 \leq i \leq n$ is $[1 - (P_{s(1)} + P_{s(A)}]^n$. As n approaches 0, this probability approaches 0.

Theorem 3 *Under the conditions of Theorem 2, for any integer $i \geq 1$, $\hat{\theta}_i - \theta_i$ does not converge in probability to 0.*

Proof. If $S_i = s(1)$, then $\hat{\theta}_i = -\infty$. If $S_i = s(A)$, then $\hat{\theta}_i = \infty$. Because the probabilities $P_{s(1)}$ and $P_{s(A)}$ defined in the proof of Theorem 2 are positive and constant and because $\hat{\theta}_i = -\infty$ with probability at least $P_{s(1)}$ and $\hat{\theta}_i = \infty$ with probability at least $P_{s(A)}$, it follows that $\hat{\theta}_i - \theta_i$ does not converge in probability to 0.

The inconsistency of $\hat{\boldsymbol{\beta}}$ is less obvious in the case of extended joint maximum-likelihood estimates. Some insight is readily provided through an examination of the statistical properties of the counts f_{sjk}. This examination can be used to show that $\hat{\boldsymbol{\beta}}$ converges with a probability of 1 to a limit $\boldsymbol{\beta}_M$ that is not necessarily $\boldsymbol{\beta}$. Demonstration of this claim requires a study of the expectation $E(f_{sjk})$ of f_{sjk}. If m_{sjkC} is the conditional expectation of Z_{ijk} given $S_i = s$, then $E(f_{sjk})$ is $nP_s m_{sjkC}$. As in the Rasch model, m_{kjC} depends on the array $\boldsymbol{\beta}$ of item parameters but not of the examinee ability θ_i. Let

$$M_{sjk}(\boldsymbol{\beta}) = \sum_{\mathbf{c} \in \Gamma(s)} \delta_k(c_j) \exp[-\boldsymbol{\beta}' \mathbf{Z}(\mathbf{c})]$$

be the partial derivative of $M_s(\boldsymbol{\beta})$ with respect to β_{jk}, where $M_s(\boldsymbol{\beta})$ is defined as in (4). The conditional probability that $\mathbf{Y}_1 = \mathbf{c}$ in $\Gamma(s)$ given $S_1 = s$ is

$$p_{JC}(\mathbf{c}) = [M_s(\boldsymbol{\beta})]^{-1} \exp[-\boldsymbol{\beta}' \mathbf{Z}(\mathbf{c})],$$

so that m_{sjkC} is

$$m_{sjk}(\boldsymbol{\beta}) = \frac{M_{sjk}(\boldsymbol{\beta})}{M_s(\boldsymbol{\beta})}.$$

Normally m_{sjk} is positive; however, $m_{sjk}(\boldsymbol{\beta})$ is 0 if $s = s(1)$ and $u_j(k) > u_{j-}$ or $s = s(A)$ and $u_j(k) < u_{j+}$. With these preliminary results, the following theorem is available.

Theorem 4 *Under the conditions of Theorem 2, $\hat{\theta}_{sC}$ converges almost surely to θ_{sM}, s in $\mathcal{S}$, and $\hat{\boldsymbol{\beta}}$ converges almost surely to $\boldsymbol{\beta}_M$, where $\theta_{s(1)M} = -\infty$, $\theta_{s(A)M} = \infty$, real θ_{sM}, s in $\mathcal{S}$, $s(1) < s < s(A)$, $\boldsymbol{\beta}_M$ in Λ, and real p_{sjkM}, s in $\mathcal{S}$, $1 \leq j \leq q$, and $1 \leq k \leq r_j$, are uniquely determined by the following conditions:*

$$p_{sjkM} = p_{jk}(\boldsymbol{\beta}_M, \theta_{sM}) \tag{15}$$

for s in $\mathcal{S}$ such that $s(1) < s < s(A)$, $p_{sjkM} = 0$ for $u_j(k) > u_{j-}$ and $s = s(1)$,

$$p_{sjkM} = e_{j1}(\boldsymbol{\beta}_M) \exp(-\beta_{jkM}) \tag{16}$$

for $s = s(1)$, $u_j(k) = u_{j-}$, β_{jkM} coordinate $\zeta(j,k)$ of $\boldsymbol{\beta}_M$, and $[e_{j1}(\boldsymbol{\beta}_M)]^{-1}$ the sum of $\exp(-\beta_{jkM})$ for k from 0 to $r_j - 1$ for which $u_j(k) = u_{j-}$, $p_{sjkM} = 0$ for $u_j(k) < u_{j+}$ and $s = s(A)$,

$$p_{sjkM} = e_{j2}(\boldsymbol{\beta}_M) \exp(-\beta_{jkM}) \tag{17}$$

for $s = s(A)$, $u_j(k) = u_{j+}$, and $[e_{j2}(\boldsymbol{\beta}_M)]^{-1}$ the sum of $\exp(-\beta_{jkM})$ for k from 0 to $r_j - 1$ for which $u_j(k) = u_{j+}$,

$$\sum_{s \in \mathcal{S}} P_s p_{sjkM} = \sum_{s \in \mathcal{S}} P_s m_{sjkC} \tag{18}$$

for $1 \leq j \leq q$ and $0 \leq k \leq r_j - 1$, and

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} u_j(k) p_{sjkM} = \sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} u_j(k) m_{sjkC} = s \tag{19}$$

for s in $\mathcal{S}$

Proof. The strong law of large numbers implies that $n^{-1} f_{sjk}$ converges almost surely to $P_s m_{sjkC}$. Existence and uniqueness of θ_{sM}, $\boldsymbol{\beta}_M$, and p_{sjkM} follow from standard results for log-linear models (Haberman, 1974, chaps. 2, 9). Results on almost sure convergence follow from general results on concave likelihood functions (Haberman, 1989).

To interpret the limit parameters θ_{sM}, $\boldsymbol{\beta}_M$, and p_{sjkM}, logarithmic penalty functions may be employed (Gilula & Haberman, 1994, 1995). Let

$$H_j(\mathbf{x}, \mathbf{y}) = -\sum_{k=0}^{r_j - 1} x_k \log(y_k)$$

for r_j-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ with respective nonnegative coordinates x_k and y_k, $0 \leq k \leq r_j - 1$, such that

$$\sum_{k=0}^{r_j-1} x_k = \sum_{k=0}^{r_j-1} y_k = 1.$$

In the definition of $H_j(\mathbf{x}, \mathbf{y})$, $0 \log 0 = 0$. Consider probability prediction of the responses $\mathbf{Y}_1$ from the sums S_1 under the incorrect model that, conditional on $S_1 = s$, s in $\mathcal{S}$, the Y_{ij}, $1 \leq j \leq q$, are independently distributed with probability

$$\pi_{sjk} = p_{jk}(\boldsymbol{\beta}_0, \theta_{s0})$$

that $Y_{ij} = k$ for unknown real parameters θ_{s0}, s in $\mathcal{S}$, and $\boldsymbol{\beta}_0$ in Λ. Let $\mathbf{m}_{sjC}$ be the r_j-dimensional vector with coordinates m_{sjkC}, and let $\boldsymbol{\pi}_{sj}$ be the r_j-dimensional vector with coordinates π_{sjk}. The expected logarithmic penalty per item is

$$q^{-1} \sum_{s \in \mathcal{S}} \sum_{j=1}^{q} P_s H_j(\mathbf{m}_{sjC}, \boldsymbol{\pi}_{sj}).$$

Let $\mathbf{p}_{sjM}$ be the r_j-dimensional vector with coordinates p_{sjkM} for $1 \leq k \leq r_j$. Then the minimum expected penalty per item is

$$H_J = q^{-1} \sum_{s \in \mathcal{S}} \sum_{j=1}^{q} P_s H(\mathbf{m}_{sjC}, \mathbf{p}_{sjM}).$$

The expected penalty per observation approaches H_J if θ_{s0} approaches θ_{sM}, $s \in \mathcal{S}$, and $\boldsymbol{\beta}_0$ approaches $\boldsymbol{\beta}_M$. Theorem 4 implies that the estimated expected log penalty function per item

$$\hat{H}_J = -\frac{1}{nq} \ell_{JCM}$$

converges almost surely to H_J.

Theorem 4 implies that inconsistency of $\hat{\boldsymbol{\beta}}$ is observed when $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_M$ differ. This situation is typically but not necessarily the case, as is evident from previous work on the binary Rasch model (Andersen, 1973a).

The expected logarithmic penalty per item H_J is at least as large as the conditional entropy measure per item

$$H_M = -q^{-1} \sum_{s \in \mathcal{S}} P_s \sum_{j=1}^{q} H_j(\mathbf{m}_{sjC}, \mathbf{m}_{sjC})$$

that corresponds to the conditional entropy per item of Y_{1A} given S_1 for a random variable A uniformly distributed on the integers 1 to q and independent of the Y_{ij}, $1 \leq j \leq q$. One has $H_J = H_M$ if, and only if, $p_{sjkM} = m_{sjkC}$. Let $\hat{m}_{sjC}$ be the r_j-dimensional vector with coordinates $\hat{m}_{sjkC}$, $1 \leq k \leq r_j$, where $\hat{m}_{sjkC} = f_{sjk}/N_S(s)$ for $N_S(s) > 0$ and $\hat{m}_{sjkC} = f_{+jk}/n$ otherwise. The entropy per item H_M has an estimate

$$\hat{H}_M = -\frac{1}{nq} \sum_{s \in \mathcal{S}} N_S(s) H_j(\hat{\mathbf{m}}_{sjC}, \hat{\mathbf{m}}_{sjC})$$

that converges almost surely to H_M.

For an R_q-dimensional vector $\mathbf{x}$, let the maximum norm $|\mathbf{x}|$ be the maximum absolute value of the coordinates of $\mathbf{x}$. As in the binary Rasch model (Haberman, 2004), the magnitude of the maximum norm $|\boldsymbol{\beta}_M - \boldsymbol{\beta}|$ is of order q^{-1}. For a formal statement and proof of this claim, consider the following theorem in which the number of items is allowed to increase.

Theorem 5 *A real number $\tau > 0$ exists such that $|\boldsymbol{\beta}_M - \boldsymbol{\beta}| < \tau/q$ for all $q \geq 1$ and all items j, $1 \leq j \leq q$, and values k, $1 \leq k \leq r_j$.*

Proof. To verify this claim, consider the difference between m_{sjkC} and

$$p_{sjkC} = p_{jk}(\boldsymbol{\beta}, \theta_{sC}),$$

where p_{sjkC} is uniquely defined by the condition that

$$\sum_{j=1}^{q} u_j(k) p_{sjC} = s$$

(Haberman, 1974, chap. 10). Let ω_{sj} be independent observations with probability p_{sjkC} that $L_{sj} = k$. The conditional probability m_{sjkC} that $Y_{ij} = k$ given that $S_i = s$ is then the conditional probability that $\omega_{sj} = k$ given that

$$L_s = \sum_{j=1}^{q} u_j(\omega_{sj}) = s.$$

This latter probability is then

$$P(\omega_{sj} = k) P\left(L_s - u_j(\omega_{sj}) = s - u_j(k)\right) / P(L_s = s).$$

Let

$$\mu_{sjC} = \sum_{k=0}^{r_j-1} u_j(k) p_{sjkC}$$

be the mean,

$$\sigma^2_{sjC} = \sum_{k=0}^{r_j-1} [u_j(k) - \mu_{sjC}]^2 p_{sjkC}$$

be the variance, and

$$\mu_{3sjC} = \sum_{k=0}^{r_j-1} [u_j(k) - \mu_{sjC}]^3 p_{sjkC}$$

be the third central moment of $u_j(K_{sj})$. If s, q, and n are selected so that

$$\sigma^2_{s+C} = \sum_{j=1}^{q} \sigma^2_{sjC}$$

approaches ∞, then

$$(L_s - s)/\sigma_{s+C}$$

and

$$[L_s - u_j(\omega_{sj}) - s + \mu_{sjC}]/(\sigma^2_{s+C} - \sigma^2_{sjC})^{1/2}$$

converge in distribution to a standard normal random variable (Cramér, 1946, pp. 215–216). A refinement of this result permits approximation of m_{sjC} (Haberman, 2004). To derive the desired approximations requires some simple modifications of results on Edgeworth expansions for lattice distributions (Esseen, 1945). Terms are used based on the normal density function and on its first three derivatives. Let

$$\psi_s = -\frac{1}{\sigma^2_{s+C}} \sum_{j=1}^{q} \mu_{3sjC},$$

so that $-\psi_k/\sigma_{k+C}$ is the skewness coefficient of L_s. Let

$$\Delta_{sjk} = \frac{\sigma^2_{sjC} - [u_j(k) - \mu_{sjC}]^2 - [u_j(k) - \mu_{sjC}]\psi_s}{2\sigma^2_{s+C}}.$$

It then follows that

$$\sigma^4_{s+C}[m_{sjkC} - p_{sjkC}(1 + \Delta_{sjk})],\ 1 \le j \le q$$

is uniformly bounded. This result indicates that $m_{sjkC} - p_{sjkC}$ is of order q^{-1}. Consider the conditional entropy H_B of Y'_B given S_1 and B for B uniformly distributed on the integers 1 to q and Y'_j random variables for $1 \leq j \leq q$ such that $P(Y'_j = k|S_1 = k) = p_{sjkC}$. Then

$$H_B = -q^{-1} \sum_{s \in \mathcal{S}} P_s \sum_{j=1}^{q} H(\mathbf{p}_{sjC}, \mathbf{p}_{sjC})$$

and H_M differ by a term of order q^{-1}.

To show that $|\boldsymbol{\beta}_M - \boldsymbol{\beta}|$ is of order q^{-1} requires use of fixed point theorems (Loomis & Sternberg, 1968, pp. 228–234). Consider solution of (18) for s in $\mathcal{S}$ subject to the constraints that $\boldsymbol{\beta}$ is in Λ and (15) and (19) hold. For an R_q-dimensional vector $\mathbf{x}$ there is a unique real value $g_s(\mathbf{x})$, s in $\mathcal{S}$, $s(1) < s < s(A)$, for which

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j - 1} u_j(k) p_{jk}(\mathbf{x}, g_s(\mathbf{x})) = s.$$

Let

$$w_j(\mathbf{x}, \theta) = \sum_{k=0}^{r_j - 1} [u_j(k)]^2 p_{jk}(\mathbf{x}, \theta) - \left[\sum_{k=0}^{r_j - 1} u_j(k) p_{jk}(\mathbf{x}, \theta) \right]^2,$$

let

$$p_{sjk}(\mathbf{x}) = p_{jk}(\mathbf{x}, g_s(\mathbf{x})),$$

let

$$\mu_{sj}(\mathbf{x}) = \sum_{k=0}^{r_j - 1} u_j(k) p_{sjk}(\mathbf{x}),$$

let

$$w_{sj}(\mathbf{x}) = w_j(\mathbf{x}, g_s(\mathbf{x})),$$

and let

$$w_{s+}(\mathbf{x}) = \sum_{j=1}^{q} w_{sj}(\mathbf{x}).$$

The function g_s is infinitely differentiable, and the partial derivative of $g_s(\mathbf{x})$ with respect to x_{jk}, the coordinate $\zeta(j,k)$ of $\mathbf{x}$, is

$$g_{sjk}(\mathbf{x}) = \frac{1}{w_{s+}(\mathbf{x})} [u_j(k) - \mu_{sj}(\mathbf{x})] p_{sjk}(\mathbf{x}).$$

For $s = s(1)$ and $u_j(k) = u_{j-}$, let

$$p_{sjk}(\mathbf{x}) = e_{j1}(\mathbf{x}) \exp(-x_{jk}),$$

and for $s = s_-$ and $u_j(k) > u_{j-}$, let

$$p_{sjk}(\mathbf{x}) = 0.$$

For $s = s(A)$ and $uk_j(k) = u_{j+}$, let

$$p_{sjk}(\mathbf{x}) = e_2(\mathbf{x}_j) \exp(-x_{jk}),$$

while for $s = s(A)$ and $u_j(k) < u_{j+}$, let

$$p_{sjk}(\mathbf{x}) = 0.$$

Let $\mathbf{F}(\mathbf{x}, \mathbf{y})$ be defined for $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}$ an R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to x_{jk}, $1 \le j \le q$, $0 \le k \le r_j - 1$, and $\mathbf{y}$ an R_qA-dimensional vector with coordinate $R_q(a-1) + \zeta(j,k)$ equal to $y_{s(a)jk}$, $1 \le a \le A$, $1 \le j \le q$, $0 \le k \le r_j - 1$, so that $\mathbf{F}(\mathbf{x}, \mathbf{y})$ is the R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to

$$F_{jk}(\mathbf{x}, \mathbf{y}) = \sum_{s \in \mathcal{S}} [y_{sj+} p_{sjk}(\mathbf{x}) - y_{sjk}]$$

for $1 \le j \le q$ and $0 \le k \le r_j - 1$. Here

$$y_{sj+} = \sum_{k=0}^{r_j - 1} y_{sjk}.$$

Then

$$\mathbf{F}(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{0}$$

for $z_{sjk} = P_s p_{sjkC}$, and

$$\mathbf{F}(\boldsymbol{\beta}_M, \mathbf{y}') = 0$$

for $\mathbf{y}'$ with $y'_{sjk} = P_s p_{sjkM}$. The conclusions of the theorem follow from application of fixed point theorems to $\mathbf{F}$. Arguments are quite similar to those previously applied to the binary Rasch model (Haberman, 2004). As a consequence, only the required derivatives are described in the remainder of the proof.

The function $\mathbf{F}(\mathbf{x}, \mathbf{y})$ of $\mathbf{x}$ and $\mathbf{y}$ is infinitely differentiable. It is linear in the second argument $\mathbf{y}$. The partial derivative of F_{jk} with respect to $x_{j'k'}$ is

$$F_{jkj'k'}(\mathbf{x}, \mathbf{y}) = \sum_{s \in \mathcal{S}} y_{sj+} F_{sjkj'k'}(\mathbf{x}),$$

where $F_{sjkj'k'}(\mathbf{x})$ is defined in the following fashion. For s in $\mathcal{S}$ and $s(1) < s < s(A)$,

$$\begin{aligned} F_{sjkj'k'}(\mathbf{x}) &= -y_{sj+} p_{sjk}(\mathbf{x}) \{\delta_j(j')[\delta_k(k') - p_{sjk'}(\mathbf{x})] \\ &\quad + [u_j(k) - \mu_{sj}(\mathbf{x})][u_{j'}(k') - \mu_{sj'}(\mathbf{x})] p_{sj'k'}(\mathbf{x}) / w_{s+}(\mathbf{x})\}. \end{aligned}$$

For $s = s(1)$ or $s = s(A)$,

$$F_{sjkj'k'}(\mathbf{x}) = -y_{sj+} p_{sjk}(\mathbf{x}) \delta_j(j')[\delta_k(k') - p_{sjk'}(\mathbf{x})].$$

Given the definitions of θ_{sM} and θ_{sC} and the properties of g_s, it also follows that $\theta_{sM} - \theta_{sC}$ and $p_{sjkM} - p_{sjkC}$ are of order q^{-1} if s/q converges to a constant greater than s^*_- and less than s^*_+. More precise expressions for these differences can be obtained but are not especially attractive.

A variety of entropy measures are closely linked. The difference $H_J - H_M$ is of order q^{-2}, so that $H_J - H_B$ is of order q^{-1}. Let $\mathbf{p}_{ij}$ be the r_j-dimensional vector with coordinates p_{ijk} for $1 \leq k \leq r_j$. With a similar argument based on the normal approximation for the distribution of S_1 given θ_1, it follows that $H_B - H_\theta$ is of order q^{-1} if

$$H_\theta = -q^{-1} \sum_{j=1}^{q} E(H_j(\mathbf{p}_{1j}, \mathbf{p}_{1j}))$$

is the conditional entropy per item of $\mathbf{Y}_1$ given θ_1.

The assumption that the numerator $u_{jn}(k)$ and denominator $u_{jd}(k)$ are uniformly bounded implies that an integer $u > 0$ exists such that $A \leq uq$ elements for each value of q. The conditional entropy per item

$$H_{+\theta} = -q^{-1} \sum_{s \in \mathcal{S}} E\left((P(S_1 = s|\theta_1) \log P(S_1 = s|\theta_1)\right)$$

of S_1 given θ_i and the unconditional entropy per item

$$H_+ = -q^{-1} \sum_{s \in \mathcal{S}} P_s \log P_s$$

of S_1 cannot exceed $q^{-1}\log(uq)$. It follows that the conditional entropy per item

$$\begin{aligned} H_C &= -q^{-1}\sum_{s\in\mathcal{S}} P_s \sum_{\mathbf{c}\in\Gamma(s)} p_{JC}(\mathbf{c})\log p_{JC}(\mathbf{c}) \\ &= H_\theta - H_{+\theta} \end{aligned}$$

of $\mathbf{Y}_1$ given S_1 and θ_1 differs from H_θ by a term of order $q^{-1}\log q$. The conditional distribution of $\mathbf{Y}_1$ given S_1 and θ_1 is assumed independent of θ_1, so that H_C is also the conditional entropy per item of $\mathbf{Y}_1$ given S_1. It follows that H_C differs from H_B, H_J, and H_M by terms of order $q^{-1}\log q$. The unconditional entropy per item

$$\begin{aligned} H_U &= -q^{-1}\sum_{s\in\mathcal{S}} \sum_{\mathbf{c}\in\Gamma(s)} p_J(\mathbf{c})\log p_J(\mathbf{c}) \\ &= H_C + H_+ \end{aligned}$$

of $\mathbf{Y}_1$ differs from H_θ, H_C, H_J, H, and H_M by terms of order $q^{-1}\log q$.

Consistency if the number of items increases. Given that the bias magnitude $|\boldsymbol{\beta}_M - \boldsymbol{\beta}|$ is reduced as q increases, there is the suggestion that the inconsistency of the joint maximum likelihood estimators for the Rasch model can be removed if the asymptotic framework is changed so that both the sample size n and the number of items q both approach infinity (Haberman, 1977, 2004). The previous argument with fixed-point theorems is easily modified. The normal approximations for the sums

$$\sum_{s\in\mathcal{S}}[f_{sjk} - N_S(s)m_{sjkC}]$$

and large-deviation arguments may be used to demonstrate that the probability approaches 1 that $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M|$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$ both converge in probability to 0.

Arguments from the binary Rasch model may be applied virtually without change to study the distribution of θ_1 (Haberman, 2004). Both $\hat{\theta}_{sM} - \theta_{sC}$ and $\hat{p}_{sjkM} - p_{sjkC}$ converge in probability to 0 if s/q converges to a constant greater than s_-^* and less than s_+^*. In turn, it follows that, for any specific individual i, $\hat{\theta}_i$ converges in probability to θ_i. Thus for any real $\delta > 0$, the fraction of examinees $i \le n$ with $|\hat{\theta}_i - \theta_i| > \delta$ converges in probability to 0. To estimate the distribution function D of the random variable θ_i, let $\hat{D}$ be the empirical distribution function of the $\hat{\theta}_i$, so that $\hat{D}(x)$ is the fraction of the $\hat{\theta}_i$ that do not exceed the

real number x. If D is continuous at x, then $|\hat{D}(x) - D(x)|$ converges in probability to 0. If h is a continuous or piecewise-continuous bounded function on the extended real line and h is continuous at θ_1 with a probability of 1, then

$$\hat{E}(h(\theta)) = n^{-1} \sum_{i=1}^{n} h(\hat{\theta}_i)$$

converges in probability to $E(h(\theta_1))$. If the distribution function D is continuous, as is the case for θ_1 a continuous random variable, then

$$|\hat{D} - D| = \sup_x |\hat{D}(x) - D(x)|$$

converges in probability to 0.

The difference $\hat{H}_J - H_U$ then converges in probability to 0, so that the various conditional entropies under study can be estimated. The difference $\hat{H}_M - H_M$ can only be expected to converge in probability to 0 if q^2/n approaches 0.

To ensure that all $\hat{\theta}_i$ are finite requires the condition that $nP_{s(1)}$ and $nP_{s(A)}$ both approach 0. This condition will certainly hold if $q^{-1} \log n$ approaches 0 (Haberman, 1977). In this case, the probability approaches 1 that

$$\max_{1 \le i \le n} |\hat{\theta}_i - \theta_i|$$

and

$$\max_{1 \le i \le n} \max_{1 \le j \le q} \max_{0 \le k \le r_j - 1} |\hat{p}_{ijk} - p_{ijk}|$$

converge in probability to 0.

For the TOEFL example, the consistency results are fairly satisfactory. The sample size of $n = 2,720$ is large enough so that the basic consistency results for $\hat{\boldsymbol{\beta}}$ are not a problem if the model is correct. Because $q = 71$ and $q^{-1} \log n$ is 0.111, there is reason for concern about the results that involve consistency of the ability estimates, for 0.111 is not that small a number. This concern is justified to the extent that an observation does exist in the sample for which the ability estimate is ∞. A further constraint exists in that $q^{-1} \log q = 0.060$ is not especially small, so that the unconditional entropy H_U is not well estimated. Results are presumably worse if the 71 items are divided into the 38 items from the reading test and the 33 items from the listening test.

2.5 Normal Approximations

The bias issues already noted in the discussion of consistency have an unusual effect on normal approximations. It is relatively easy to find a normal approximation for the joint maximum-likelihood estimate $\hat{\beta}_{jk}$ of the item parameter β_{jk}, but this approximation is often not satisfactory because the asymptotic mean is β_{jkM}, coordinate $\zeta(j,k)$ of $\boldsymbol{\beta}_M$, rather than β_{jk}. A normal approximation for $\hat{\theta}_i$ is available with relatively little difficulty for q large, but there are problems in practice with the accuracy achieved. Results are rather straightforward generalizations of those for the binary case (Haberman, 2004).

If q is constant and n becomes large, then a normal approximation is available for $\hat{\beta}_{jk}$ but not for $\hat{\theta}_i$. The normal approximation is derived by conventional arguments based on the function $\mathbf{F}$ developed in Section 2.4. Once again, fixed point theorems are employed as in the binary Rasch model (Haberman, 2004). Let Z_{ijk}^+ be the adjusted random variable with value $Z_{ijk} - p_{sjkM}$ for $S_i = s$. Let $\mathbf{V}^+$ be the covariance matrix of the R_q-dimensional vector $\mathbf{Z}_i^+$ with coordinate $\zeta(j,k)$ equal to Z_{ijk}^+ for $1 \le j \le q$ and $0 \le k \le r_j - 1$. Let $V_{jkj'k'}^+$ be the covariance of Z_{ijk}^+ and $Z_{ij'k'}^+$. Let

$$\mu_{sjM} = \sum_{k=0}^{r_j-1} u_j(k) p_{sjkM}$$

and

$$\sigma_{sjM}^2 = \sum_{k=0}^{r_j-1} [u_j(k) - \mu_{sjM}]^2 p_{sjkM}$$

be the variance of a random variable that is $u_j(k)$ with probability p_{sjkM}, let

$$\sigma_{+jM}^2 = \sum_{s \in \mathcal{S}} P_s \sigma_{sjM}^2,$$

$$\sigma_{s+M}^2 = \sum_{j=1}^{q} \sigma_{sjM}^2,$$

let $\mathcal{S}'$ be the set of s in $\mathcal{S}$ that are neither $s(1)$ nor $s(A)$, let

$$T_{1jkj'k'} = \sum_{s \in \mathcal{S}} P_s \delta_j(j') p_{sjkM} [\delta_k(k') - p_{sj'k'M}],$$

let

$$T_{2jkj'k'} = \sum_{s \in \mathcal{S}'} P_s \frac{p_{sjkM} p_{sj'k'M} [u_j(k) - \mu_{sjM}][u_{j'}(k') - \mu_{sj'k'M}]}{\sigma_{s+M}^2},$$

and let

$$W_{jkj'k'} = T_{1jkj'k'} - T_{2jkj'k'}.$$

Let $\mathbf{W}$ be the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to $W_{jkj'k'}$. Note that

$$\sum_{k'=0}^{r_j-1} W_{jkj'k'} = \sum_{k'=0}^{r_j-1} V^+_{jkj'k'} = 0$$

and

$$\sum_{j'=1}^{q}\sum_{k'=0}^{r_j-1} u_{j'}(k')W_{jkj'k'} = \sum_{j'=1}^{q}\sum_{k'=0}^{r_j-1} u_{j'}(k')V^+_{jkj'k'} = 0$$

and $\mathbf{W}$ and $\mathbf{V}^+$ are symmetric and positive semi-definite. Let $\mathbf{W}^+$ be the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$W^+_{jkj'k'} = W_{jkj'k'} + \delta_j(j') + \delta_1(j)\delta_1(j')[u_1(k) - \bar{u}_1][u_1(k') - \bar{u}_1].$$

Then $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $(\mathbf{W}^+)^{-1}\mathbf{V}^+(\mathbf{W}^+)^{-1}$. The notable problem is that the normal approximation involves $\boldsymbol{\beta}_M$ rather than $\boldsymbol{\beta}$.

If the number q of items increases, then normal approximations remain available, but a few changes in results are needed due to the changing dimension of $\hat{\boldsymbol{\beta}}$. Let

$$v_{sjkj'k'}(\boldsymbol{\beta}) = \frac{M_{sjkj'k'}(\boldsymbol{\beta})}{M_s(\boldsymbol{\beta})} - m_{sjk}(\boldsymbol{\beta})m_{sj'k'}(\boldsymbol{\beta}),$$

where

$$M_{sjkj'k'}(\boldsymbol{\beta}) = \sum_{\mathbf{c}\in\Gamma(s)} c_j c_{j'} \exp[-\boldsymbol{\beta}'\mathbf{Z}(\mathbf{c})].$$

Thus $v_{sjkj'k'}(\boldsymbol{\beta})$ is the conditional covariance of Z_{1jk} and $Z_{1j'k'}$ given $S_1 = s$. Arguments similar to those applied for m_{sjkC} may be used to show that

$$\sigma^4_{s+C}[v_{sjkjk}(\boldsymbol{\beta}) - \sigma^2_{kjC} - (2p_{sjkC} - 1)p_{sjkC}\Delta_{sjk}],\ 1 \le j \le q, 0 \le k \le r_j - 1,$$

$$\sigma^4_{s+C}[v_{sjkjk'}(\boldsymbol{\beta}) + p_{sjkC}p_{sjk'C}(1 + \Delta_{sjk} + \Delta_{sjk'})],\ 1 \le j \le q, 0 \le k < k' \le r_j - 1,$$

and

$$\begin{aligned}\sigma^4_{s+C}\{v_{sjkj'k'}(\boldsymbol{\beta}) \quad &+ \quad p_{sjkC}p_{sj'k'C}[u_j(k) - \mu_{sjkC}][u_{j'}(k') - \mu_{sj'k'C}]/\sigma^2_{s+C}\},\\ &1 \le j < j' \le q, 0 \le k \le r_j - 1, 0 \le k' \le r_{j'} - 1,\end{aligned}$$

are uniformly bounded as σ^2_{s+C} approaches ∞.

Let Q be an integer constant greater than 1. For $q \geq Q$, let $\hat{\boldsymbol{\beta}}_Q$ be the R_Q-dimensional vector with coordinate $\zeta(j,k)$ of $\hat{\beta}_{jk}$ for $1 \leq j \leq Q$ and $0 \leq k \leq r_j - 1$, let $\boldsymbol{\beta}_{QM}$ be the R_Q-dimensional vector with coordinate $\zeta(j,k)$ of β_{jkM} for $1 \leq j \leq Q$ and $0 \leq k \leq r_j - 1$, and let $\boldsymbol{\beta}_Q$ be the R_Q-dimensional vector with coordinate $\zeta(j,k)$ equal to β_{jk} for $1 \leq j \leq Q$ and $0 \leq k \leq r_j - 1$. Let $\mathbf{T}_Q$ be the R_Q by R_Q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$T_{jkj'k'} = E(\delta_j(j')p_{ijk}[\delta_k(k') - p_{ij'k'}]).$$

Let $\mathbf{T}^+_Q$ be the R_Q by R_Q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$T^+_{jkj'k'} = \delta_j(j') + T_{jkj'k'}.$$

Let $\mathbf{K}_Q$ be the R_Q by R_Q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$K_{jkj'k'} = \delta_j(j')\delta_k(k') - U_1^{-1}\delta_1(j')u_{jk}u_{j'k'}.$$

Arguments can be used similar to those for the binary Rasch model (Haberman, 2004). Use of the maximum norm shows that $n^{1/2}(\hat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_{QM})$ converges in distribution to a multivariate normal random vector with zero mean and covariance matrix $\mathbf{K}_Q(\mathbf{T}^+_Q)^{-1}\mathbf{T}_Q(\mathbf{T}^+_Q)^{-1}\mathbf{K}'_Q$, where the prime symbol is used to denote a transpose.

In practice, the asymptotic normality result is somewhat unsatisfactory. Clearly $\hat{\beta}_{jk}$ is intended to estimate β_{jk} rather than β_{jkM}. If n/q^2 approaches 0, then $n^{1/2}(\hat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_Q)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}_Q(\mathbf{T}^+_Q)^{-1}\mathbf{T}_Q(\mathbf{T}^+_Q)^{-1}\mathbf{K}'_Q$. Nonetheless, it is far from clear that the asymptotic approximation is adequate for the example under study, for n/q^2 is 0.540 is not a small number. The problem is likely to be much more severe with tests of similar length in which an administration may involve around 500,000 examinees. As a practical matter, the results indicate that ordinary asymptotic confidence intervals for β_{jk} cannot be derived by use of the normal approximation for $\hat{\beta}_{jk}$.

In the case of an individual i for an increasing number q of items, the normal approximation for $\hat{\theta}_i$ is relatively straightforward. Arguments for the binary Rasch model apply with only minor modifications (Haberman, 2004). One finds that $(\hat{\theta}_i - \theta_i)/\sigma(\hat{\theta}_i)$

converges in distribution to a standard normal random variable if $\sigma(\hat{\theta}_i)$ is the inverse of

$$\sum_{j=1}^{q} \sigma_{ij}^2,$$

$$\sigma_{ij}^2 = \sum_{k=1}^{r_j} p_{ijk}[u_j(k) - \mu_{ij}]^2,$$

and

$$\mu_{ij} = \sum_{k=1}^{r_j} u_j(k) p_{ijk}.$$

In addition, for Q a finite integer, the Q-dimensional vector with coordinates $(\hat{\theta}_i - \theta_i)/\sigma(\hat{\theta}_i)$ for $1 \leq i \leq Q$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$.

Approximate confidence intervals are available. The probability that

$$\hat{\theta}_i - z\hat{\sigma}(\hat{\theta}_i) < \theta_i < \hat{\theta}_i + z\hat{\sigma}(\hat{\theta}_i)$$

approaches $1 - \alpha$ if

$$\hat{\sigma}(\theta_i) = 1/\hat{\sigma}_{i+},$$

$$\hat{\sigma}_{i+}^2 = \sum_{j=1}^{q} \hat{\sigma}_{ij}^2,$$

$$\hat{\sigma}_{ij}^2 = \sum_{k=1}^{r_j} \hat{p}_{ijk}[u_j(k) - \hat{\mu}_{ij}]^2,$$

and

$$\hat{\mu}_{ij} = \sum_{k=1}^{r_j} u_j(k) \hat{p}_{ijk}.$$

For the TOEFL example, the estimated $\hat{\theta}_i$ range from -2.632 to 4.326 for the cases with finite estimates. One value of $\hat{\theta}_i$ is ∞. The observed estimates $\hat{\sigma}(\hat{\theta}_i)$ range from 0.248 for scores S_i from 38 to 41 to 1.009 for $S_i = 78$. The value of $\hat{\sigma}(\hat{\theta}_i)$ is taken to be ∞ for $S_i = 79$. The lower quartile for the $\hat{\theta}_i$ is -0.448, and the upper quartile is 1.164. The estimates of asymptotic standard deviations suggest some limitations in the quality of the normal approximations.

Normal approximations for $\hat{H}_J$ and $\hat{H}_M$ are somewhat unsatisfactory in practice due to the relatively large estimation biases involved.

3. Conditional Maximum Likelihood

Conditional maximum-likelihood estimation is applicable to the nominal model (Andersen, 1983), and conditional maximum-likelihood is closely related to marginal maximum likelihood. As shown in this section, conditional maximum likelihood is quite effective in large samples whether or not the number of items is large, and computation of conditional maximum-likelihood estimates is relatively straightforward. In conditional maximum likelihood, inference is conditional on the observed examinee sums S_i. For $\mathbf{c}$ in $\Gamma(s)$ and for s in $\mathcal{S}$, the conditional probability $p_{JC}(\mathbf{c})$ that $\mathbf{Y}_i = \mathbf{c}$ given that $S_i = s$ satisfies

$$p_{JC}(\mathbf{c}) = p_J(\mathbf{c})/P_s.$$

Under the nominal model,

$$P_s = M_s(\boldsymbol{\beta}) \exp(\tau_s),$$

so that

$$p_{JC}(\mathbf{c}) = \exp(-\boldsymbol{\beta}'\mathbf{c})/M_s(\boldsymbol{\beta}) \tag{20}$$

does not depend on the distribution function D of the ability θ_1. The conditional log likelihood function is then

$$\ell_C(\mathbf{p}_{JC}) = \sum_{i=1}^{n} \log p_{JC}(\mathbf{Y}_i)$$

for the array $\mathbf{p}_{JC}$ of $p_{JC}(\mathbf{c})$ for $\mathbf{c}$ in Γ. Thus

$$\ell_C(\mathbf{p}_{JC}) = -\boldsymbol{\beta}'\mathbf{Z}_+ - \sum_{s\in\mathcal{S}} n_s \log M_s(\boldsymbol{\beta}).$$

Because ℓ_C is determined by the f_{sjk}, inferences again may be based on the collapsed table.

As in the binary Rasch model (Haberman, 2004), the relationship of conditional and marginal maximum likelihood is relatively simple. Let $\mathbf{P}$ be the array with coordinates P_s for s in $\mathcal{S}$, and let

$$\ell_S(\mathbf{P}) = \sum_{s\in\mathcal{S}} N_S(s) \log P_s$$

be the marginal log likelihood for the examinee totals S_i, $1 \le i \le n$, under the unrestricted model that $S_i = s$ with probability P_s for some nonnegative P_s such that $\sum_{s\in\mathcal{S}} P_s = 1$. Then

$$\ell(\mathbf{p}_J) = \ell_C(\mathbf{p}_{JC}) + \ell_S(\mathbf{P}).$$

Let ℓ_M denote the maximum of the log likelihood $\ell(\mathbf{p}_J)$ under the condition that $\mathbf{p}_J$ is in the set Ξ corresponding to the nominal model, and let ℓ_{M+} denote the maximum of $\ell(\mathbf{p}_J)$ under the assumption that $\mathbf{p}_J$ is in the set Ξ_+ that corresponds to the log-linear extension of the nominal model. Obviously $\ell_M \leq \ell_{M+}$. Let ℓ_{CM} be the maximum of $\ell_C(\mathbf{p}_{JC})$ under the constraint that (20) holds for some $\boldsymbol{\beta}$ in Λ. Let ℓ_{SM} be the maximum

$$\sum_{s \in \mathcal{S}} N_S(s) \log[N_S(s)/n]$$

of $\ell_S(\mathbf{P})$ ($0 \log 0$ is taken to be 0). Then

$$\ell_M \leq \ell_{M+} = \ell_{CM} + \ell_{SM}.$$

Thus conditional maximum likelihood corresponds to ordinary maximum likelihood for the log-linear extension of the nominal model.

The conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_*$ of $\boldsymbol{\beta}$, if it exists, is the element of Λ such that

$$\hat{p}_{JC}(\mathbf{c}) = \exp(-\hat{\boldsymbol{\beta}}_*' \mathbf{c})/M_s(\hat{\boldsymbol{\beta}}_*),$$

and

$$\ell_C(\hat{\mathbf{p}}_{JC}) = \ell_{CM}.$$

If $\hat{\boldsymbol{\beta}}_*$ exists, then it satisfies the conditional maximum-likelihood equations

$$\hat{m}_{sjkC} = m_{sjk}(\hat{\boldsymbol{\beta}}_*)$$

and

$$\sum_{s \in \mathcal{S}} N_S(s) \hat{m}_{sjkC} = Z_{+jk}$$

for $1 \leq j \leq q$ and $0 \leq k \leq r_j - 1$. Conversely, if $\boldsymbol{\beta}_*$ is a vector in Λ and

$$\sum_{s \in \mathcal{S}} s N_S(s) m_{sjk}(\boldsymbol{\beta}_*) = Z_{+jk},$$

then $\boldsymbol{\beta}_*$ is a conditional maximum-likelihood estimate of $\boldsymbol{\beta}$. Provided that $\mathcal{S}'$ is nonempty, no more than one conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_*$ exists.

Existence of conditional maximum-likelihood estimates is an issue, although normally a much less important one than in the case of joint estimation. Consider the following theorem (Haberman, 1974, chaps. 2, 7)

Theorem 6 *In the case of $\mathcal{S}'$ nonempty, the estimate $\hat{\boldsymbol{\beta}}_*$ exists if, and only if, $g_{sjk} \leq 0$ can be found for s in $\mathcal{S}_+$, $0 \leq k \leq r_j - 1$, and $1 \leq j \leq q$, such that $g_{sjk} > 0$ for s in $\mathcal{S}_+$ if* **c** *in $\Gamma(s)$ exists such that $c_j = k$, $g_{sjk} = 0$ otherwise, $g_{+jk} = f_{+jk}$ for $0 \leq k \leq r_j - 1$, and $1 \leq j \leq q$, and $\sum_{j=1}^{q} \sum_{k=0}^{r_j-1} g_{sjk} = qN_S(s)$ for s in $\mathcal{S}_+$.*

It follows that conditional maximum-likelihood estimates exist whenever joint maximum-likelihood estimates exist.

Extended conditional maximum-likelihood estimates may be considered if $\hat{\boldsymbol{\beta}}_*$ does not exist. There are $\hat{m}_{sjkC}$ in $[0, 1]$ such that

$$\sum_{s \in \mathcal{S}} N_S(s) \hat{m}_{sjkC} = Z_{+jk},$$

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j-1} u_j(k) \hat{m}_{sjkC} = s,$$

and $m_{sjk}(\boldsymbol{\beta})$ approaches $\hat{m}_{sjkC}$ for $N_S(s) > 0$ if (20) holds and $\ell_C(\mathbf{p}_{JC})$ approaches ℓ_{CM}. If the conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_*$ exists, then $\hat{m}_{sjkC} = m_{sjk}(\hat{\boldsymbol{\beta}}_*)$. Various conventions can be considered to define $\hat{\boldsymbol{\beta}}_*$ in the case in which no conditional maximum-likelihood estimate exists for $\boldsymbol{\beta}$.

Given the estimate $\hat{\boldsymbol{\beta}}_*$, it is possible to estimate the examinee abilities θ_i. For each i, the log likelihood for θ_i given the $\hat{\beta}_{jk*}$ is

$$\sum_{j=1}^{q} \sum_{k=0}^{r_j-1} Z_{ijk} \log p_{jk}(\hat{\boldsymbol{\beta}}_*, \theta_i).$$

Given the definition of g_s in the proof of Theorem 5, it follows that the estimate $\hat{\theta}_{i*}$ of θ_i is

$$\hat{\theta}_{sC*} = g_s(\hat{\boldsymbol{\beta}}_*)$$

for $s(1) < s = S_i < s(A)$. For $S_i = s = s(1)$, $\hat{\theta}_{i*} = \hat{\theta}_{sC*} = -\infty$. For $S_i = s(A)$, $\hat{\theta}_{i*} = \hat{\theta}_{sC*} = \infty$.

3.1 Large-Sample Properties

For q fixed, if the nominal model is valid and n becomes large, then there is no difficulty in demonstrating that $\hat{\boldsymbol{\beta}}_C$ is a consistent and asymptotically normal estimate for

$\boldsymbol{\beta}$ (Haberman, 1977). In the case of q increasing, a bit more argument is required. Consider integers j, j', k, and k' such that $0 \le k \le r_j - 1$, $0 \le k' \le r_{j'} - 1$, $1 \le j \le q$, and $1 \le j' \le q$. Let

$$V_{jkj'k'C} = V_{jkj'k'}(\boldsymbol{\beta}) = \sum_{s \in \mathcal{S}} N_S(s) v_{sjkj'k'}(\boldsymbol{\beta})$$

be the conditional covariance of Z_{+jk} and $Z_{+j'k'}$ given the $N_S(s)$, s in $\mathcal{S}$. Let $\mathbf{V}_C = \mathbf{V}(\boldsymbol{\beta})$ be the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to $V_{jkj'k'}(\boldsymbol{\beta})$. This matrix is of rank $R_q - q - 1$ if $\mathcal{S}'$ is nonempty. Let $\mathbf{V}_C^*$ be the expected value of $n^{-1}\mathbf{V}_C$, so that $\mathbf{V}_C^*$ is obtained from $\mathbf{V}_C$ by substitution of P_s for $N_S(s)$. Note that if Z_{jk}^* is the random variable equal to $Z_{ijk} - m_{sjkC}$ for $S_i = s$ and if $\mathbf{Z}_i^*$ is the R_Q-dimensional vector with coordinate $\zeta(j,k)$ equal to Z_{ijk}^* for $1 \le j \le q$ and $0 \le k \le r_j - 1$, then $\mathbf{V}_C^*$ is the covariance matrix of $\mathbf{Z}_i^*$ for each observation i. Let $\mathbf{V}_C^+$ be the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$V_{jkj'k'C}^+ = V_{jkj'k'C}^* + \delta_j(j') + \delta_1(j)\delta_1(j')[u_1(k) - \bar{u}_1][u_1(k') - \bar{u}_1].$$

Arguments rather similar to those applied in the case of joint maximum-likelihood estimation may also be applied to conditional maximum-likelihood estimation. If the number q of items is fixed, then $\hat{\boldsymbol{\beta}}_*$ converges almost surely to $\boldsymbol{\beta}$ and $n^{1/2}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})$ converges in distribution to a multivariate normal random variable with mean $\mathbf{0}$ and covariance matrix $(\mathbf{V}_C^+)^{-1}\mathbf{V}_C^*(\mathbf{V}_C^+)^{-1}$. If q approaches ∞, then $|\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta}|$ converges in probability to 0. For an integer $Q \ge 1$, let $\hat{\beta}_{jk*}$ be coordinate $\zeta(j,k)$ of $\hat{\boldsymbol{\beta}}_*$ and let $\hat{\boldsymbol{\beta}}_{Q*}$ be the R_Q-dimensional vector with coordinate $\zeta(j,k)$ equal to $\hat{\beta}_{jk*}$ for $1 \le j \le Q$ and $0 \le k \le r_j - 1$. Then $n^{1/2}(\hat{\boldsymbol{\beta}}_{Q*} - \boldsymbol{\beta}_Q)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and the covariance matrix $\mathbf{K}_Q(\mathbf{T}_Q^+)^{-1}\mathbf{T}_Q(\mathbf{T}_Q^+)^{-1}\mathbf{K}_Q'$ encountered in the discussion of the normal approximation for $n^{1/2}(\hat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}_{QM})$. As in the binary Rasch model, conditional estimation has the major advantage that the asymptotic approximations involve the actual parameters of interest, namely the β_{jk}, rather than the β_{jkM} parameters. It should be noted that $\mathbf{K}_Q(\mathbf{T}_Q^+)^{-1}\mathbf{T}_Q(\mathbf{T}_Q^+)^{-1}\mathbf{K}_Q'$ is the limit of the matrix formed from the first R_Q rows and columns of $(\mathbf{V}_C^+)^{-1}\mathbf{V}_C^*(\mathbf{V}_C^+)^{-1}$. Let $\hat{\mathbf{V}}_C$ be $\mathbf{V}(\hat{\boldsymbol{\beta}}_*)$, and let $\hat{\mathbf{V}}_C^+$ be the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$\hat{V}_{jkj'k'C}^+ = V_{jkj'k'C}(\hat{\boldsymbol{\beta}}_*) + n\delta_j(j') + n\delta_1(j)\delta_1(j')[u_1(k) - \bar{u}_1][u_1(k') - \bar{u}_1]$$

for integers j, j', k, and k' such that $0 \leq k \leq r_j - 1$, $0 \leq k' \leq r_{j'} - 1$, $1 \leq j \leq q$, and $1 \leq j' \leq q$. Then both for q fixed and q increasing, asymptotic confidence intervals for parameters such as β_{jk} are easily constructed by estimation of the asymptotic standard deviation $s(\hat{\beta}_{jk*})$ of $\hat{\beta}_{jk*}$ by the square root of row $\zeta(j,k)$ and column $\zeta(j,k)$ of $(\hat{\mathbf{V}}_C^+)^{-1}\hat{\mathbf{V}}_C(\hat{\mathbf{V}}_C^+)^{-1}$. Thus results are quite similar to those for the binary Rasch model (Haberman, 2004).

If q approaches ∞, then the asymptotic properties of $\hat{\theta}_{i*}$ are essentially the same as those for $\hat{\theta}_i$ as far as consistency, asymptotic normality, and approximate confidence intervals are concerned. Estimation of the distribution of θ_1 can be implemented in essentially the same fashion as in JMLE by substitution of $\hat{\theta}_{i*}$ for $\hat{\theta}_i$.

Estimation of the entropy measures H_C and H_U involves relatively little difficulty, for H_C may be estimated by

$$\hat{H}_{CN} = -\frac{1}{nq}\ell_{CM},$$

H_+ may be estimated by

$$\hat{H}_+ = -\frac{1}{nq}\sum_{s\in\mathcal{S}} N_S(s)\log[N_S(s)/n],$$

and H_U may be estimated by

$$\hat{H}_{UN} = \hat{H}_{CN} + \hat{H}_+.$$

For q constant, $\hat{H}_{CN}$ converges almost surely to H_C, $\hat{H}_+$ converges almost surely to H_+, and $\hat{H}_{UN}$ converges almost surely to H_U. For q increasing, $\hat{H}_{CN} - H_C$, $\hat{H}_+ - H_+$, and $\hat{H}_{UN} - H_U$ all converge in probability to 0. Normal approximations are readily available, at least if q/n approaches 0. Let $\sigma(\hat{H}_U)$ be the standard deviation of $q^{-1}\log p_J(\mathbf{Y}_1)$, and let $\sigma(\hat{H}_C)$ be the standard deviation of $q^{-1}\log p_{JC}(\mathbf{Y}_1)$. Let $\sigma(\hat{H}_C)$ and $\sigma(\hat{H}_U)$ be positive, and assume that neither approaches 0 if q approaches ∞. Then $n^{1/2}(\hat{H}_{UN} - H_U)/\sigma(\hat{H}_U)$ and $n^{1/2}(\hat{H}_{CN} - H_C)/\sigma(\hat{H}_C)$ both converges in distribution to a standard normal random variable. These results are readily applied to construction of approximate confidence intervals for H_C and H_U (Gilula & Haberman, 1995).

3.2 The Newton-Raphson Algorithm

The Newton-Raphson algorithm for conditional estimation for the nominal model is rather similar to the Newton-Raphson algorithm for conditional estimation for the binary Rasch model (Andersen, 1972, 1983; Haberman, 2004). One begins with a preliminary approximation $\boldsymbol{\beta}_0$ to $\hat{\boldsymbol{\beta}}_*$. One possibility is $\hat{\boldsymbol{\beta}}$. One then uses the iterations

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - (\mathbf{V}_t^+)^{-1}[\mathbf{Z}_+ - \mathbf{m}_{+t}],$$

where $\mathbf{m}_{+t}$ is the R_q-dimensional vector with coordinate $\zeta(j,k)$ equal to

$$m_{+jk}(\boldsymbol{\beta}_t) = \sum_{s \in \mathcal{S}} n_s m_{sjkC}(\boldsymbol{\beta}_t)$$

for $0 \le k \le r_j - 1$ and $1 \le j \le q$ and $\mathbf{V}_t^+$ is the R_q by R_q matrix with row $\zeta(j,k)$ and column $\zeta(j',k')$ equal to

$$V^+_{jkj'k't} = V_{jkj'k'}(\boldsymbol{\beta}_t) + n\delta_j(j') + n\delta_1(j)\delta_1(j')[u_1(k) - \bar{u}_1][u_1(k') - \bar{u}_1]$$

for integers j, j', k, and k' such that $0 \le k \le r_j - 1$, $0 \le k' \le r_{j'} - 1$, $1 \le j \le q$, and $1 \le j' \le q'$. In typical cases, $\boldsymbol{\beta}_t$ converges quite rapidly to $\hat{\boldsymbol{\beta}}_*$.

Even more than for the Newton-Raphson algorithm for the binary Rasch model, implementation of the Newton-Raphson algorithm is challenging for a large number q of items. For efficient computation, consider random variables ω_{jt} and L_t defined so that the ω_{jt} are independent for $1 \le j \le q$, $L_t = \sum_{j=1}^q \omega_{jt}$; ω_{jt} assumes integer values from 0 to $r_j - 1$, and $\omega_{jt} = k$ with probability

$$p_{jkt} = p_{jk}(\boldsymbol{\beta}_{jt}, 0).$$

As in the proof of Theorem 5,

$$m_{sjkt} = m_{sjk}(\boldsymbol{\beta}_t) = p_{jkt} P(L_t - u_j(\omega_{jt}) = s - u_j(k))/P(L_t = s).$$

Similarly, $v_{sjkj'k't} = v_{sjkj'k'}(\boldsymbol{\beta}_t)$ satisfies

$$v_{sjkjkt} = m_{sjkt}(1 - m_{sjkt}),$$

$$v_{sjkjk't} = -m_{sjkt} m_{sjk't},\ k \ne k',$$

and

$$v_{sjkj'k't} = \frac{p_{sjkt}p_{sj'k't}P(L_t - u_j(\omega_{jt}) - u_{j'}(\omega_{j't}) = s - u_j(k) - u_{j'}(k')}{P(L_t = s)} - m_{sjkt}m_{sj'k't},\ j \neq j'.$$

At this point, probabilities such as $P(L_t = s)$ may be computed by use of a recursion formula. Let $\mathcal{S}_i$ be the set of possible sums $\sum_{j=1}^{i} u_j(k)$ for $0 \leq k \leq r_j - 1$ for $1 \leq j \leq i$. Let $\mathcal{S}_0$ be the set with element 0. Let $a_t(h, i)$ be the probability that $\sum_{j=1}^{i} \omega_{jt} = h$ for h in $\mathcal{S}_i$ and $1 \leq i \leq q$, and let $a(s, 0, 0) = 1$. For h in $\mathcal{S}_i$ and $1 \leq i \leq q$, let $K(h, i)$ be the set of integers k from 1 to r_i such that $h - u_i(k)$ is in $\mathcal{S}_{i-1}$. Then

$$a_t(h, i) = \sum_{k \in K(h,i)} p_{jkt} a_t(h - u_i(k), i - 1),$$

and $P(L_t = s)$ is $a_t(s, q)$.

Given that this recursion procedure is employed with double precision arithmetic, no major computational problems are encountered. The initial values from joint estimation are quite effective as starting values, for $\hat{\beta}_{jk}$ and $\hat{\beta}_{jk*}$ have no difference that exceed 0.06 in magnitude for the data from Form A of the TOEFL field trial, and most differences are much smaller in magnitude. The estimated asymptotic standard deviations range roughly from 0.03 to 0.10, so that the differences between $\hat{\beta}_{jk}$ and $\hat{\beta}_{jk*}$ can be large enough to raise some questions about the quality of large-sample approximations for JMLE.

4. Conclusions

The results derived in the preceding sections suggest that CMLE provides an effective approach for analysis of the nominal model even in cases in which both the sample size and the number of items are large. Standard large-sample approximations for the distributions of conditional maximum-likelihood estimates have been shown to apply, so that asymptotic confidence intervals are available.

Efforts have also been made to apply JMLE under realistic conditions. Results have been somewhat less satisfactory for the TOEFL example.

This report does not treat all important issues for the nominal model. Goodness of fit is an issue, and behavior of estimates when the model fails is important. The measurement of the size model error should also be explored. Generalizations of the model that are similar

to 2PL models are of interest, and use of restricted ability distributions can be explored as in conventional applications of marginal maximum likelihood.

References

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 42–54.

Andersen, E. B. (1973a). *Conditional inference and models for measuring*. Copenhagen, Denmark: Mentalhygiejnisk Forskningsinstitut.

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Andersen, E. B. (1983). A general latent structure model for contingency table data. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 117–139). Mahwah, NJ: Lawrence Elbaum Associates.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, *37*, 29–51.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*, 129–141.

Esseen, C.-G. (1945). Fourier analysis of distribution functions. a mathematical study of the Laplace-Gaussian law. *Acta Mathematica*, *77*, 1–125.

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. 2). New York: John Wiley.

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika*, *46*, 59–77.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Gilula, Z., & Haberman, S. J. (2000). Density approximation by summary statistics: An information-theoretic approach. *Scandinavian Journal of Statistics*, *27*, 521–534.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*, 815–841.

Haberman, S. J. (1989). Concavity and estimation. *The Annals of Statistics*, *17*, 1631–1661.

Haberman, S. J. (2004). *Maximum likelihood for the Rasch model for binary responses* (ETS RR-04-20). Princeton, NJ: ETS.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, *27*, 887–906.

Loomis, L. A., & Sternberg, S. (1968). *Advanced calculus.* Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.